Improving the Precision and Application of Speech Diagnostic Tests

A Dissertation

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Tzu-Ling Jocelyn Yu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Robert S. Schlauch, Ph.D.

November 2018

©Tzu-Ling Jocelyn Yu, 2018

Acknowledgements

There are no words to describe how much I thank my advisor, Bert Schlauch. You have shown me tremendous patience and acceptance in helping me walk through the transition of my program. Thank you for paving the way for my research. Your mentorship and encouragement have nourished me to grow in speech and hearing sciences.

Thank you also to my other committee members, Ben Munson, Peggy Nelson, and Pani Kendeou, for your direction, kindness, and enthusiasm throughout these years. This work is stronger because of you and your expertise you were willing to share with me.

I owe immeasurable thanks to Edward Carney, Yang Zhang, and Bert Schlauch (again), who taught me the beauty of programming languages. Your instructions empower me to do research. Thank you, Edward Carney, for your gracious patience in mentoring me in learning R and MATLAB.

I'm also very grateful to Mary Kennedy and Mark DeRuiter. Thank you for the guidance and encouragement throughout my journey of learning. You have been missed.

Many thanks in particular to my undergraduate assistants, Sydney Murray and Sarah Rosen. Each of you made crucial contributions to this project, from helping recruit participants to scoring assessments. Your cheery spirits have made this project a joy to work on.

A special thanks to the Center for Applied & Translational Sensory Science for their technical support for my data collection. I would also like to express gratitude to the

faculty and staff of Shevlin Hall, who have all helped me in too many ways over the years.

I would also like to recognize members in Schlauch lab, Heekyung Han and Andrew Kersten, and other doctoral students in SLHS, for your friendship.

To my parents: thank you for trusting me and letting me have this 10+ years-long adventure studying abroad since 2007. Your wisdom and love have been the nutrients for me to grow during these years. Thank you for giving me life and the courage to live in this world.

Finally, I want to express my profound thanks to Jeffery. I thank God for marrying you and being loved by you. You have shown me unconditional and sacrificial love for me. You left your career in Taiwan to accompany me since 2015. You had to start over and learn English in your mid-30s. It was difficult for me to fathom the depth of Jesus' love until being loved by you. Without your company and love, I would not be complete.

This dissertation project was made possible by funding from the Bryng Bryngelson Research Fund, the Doctoral Dissertation Fellowship at the University of Minnesota, and the Governmental Research Scholarship for Students Studying Abroad from the Ministry of Education in Taiwan.

Dedication

This dissertation is dedicated to Jesus Christ, showing me how much He loves me throughout this 8-year journey.

Abstract

Background: Diagnostic speech recognition tests are the most direct way to quantify the distortion component of hearing loss and to evaluate the outcome of hearing prostheses. **Purpose:** The primary purpose of this dissertation was to evaluate the diagnostic precision of the spoken word recognition (WR) tasks that differed in listeners' response formats (the closed- and open-set tasks). The second purpose was to improve the precision through a refined analysis of WR performance where the chance performance for listening parts (phonemes) of a word was considered. **Method:** WR performance for closed- and open-set tasks was obtained from seventy listeners with normal hearing. Hearing loss was simulated by presenting words in noise or in a sinewave vocoder condition. The percentage of correct phonemes in response word for each test word was computed to derive the distribution of chance performance based on an assessment of 15,000 iterations of the randomly paired response and test words. **Results:** Analyses found the following for the most to least precise and efficient conditions in detecting a change in hearing: open-set task scored by percent correct phonemes, open-set task score by percent correct words, 6-alternative closed-set task, and 4-alternative closed-set task. When the range of phoneme chance performance was accounted for in an open-set WR task, listeners with identical word scores were found to have different abilities to perceive phonemes. **Conclusions:** Closed-set WR testing has distinct advantages for implementation but its poorer precision for identifying a change in hearing than open-set WR testing must be considered. The analysis of scoring WR by phonemes on an open-set task with the estimates of chance performance reveals meaningful differences in perception that are not possible based on word scores.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

Speech recognition tests are an important measure for diagnosing hearing loss, evaluating communication equipment, and understanding the effectiveness of interventions for hearing loss, including the evaluation of prosthetic devices.

Performance of speech recognition has been found affected by differences in talker's voice features (e.g., Hood & Poole, 1980), testing environment (e.g., Neuman, Wroblewski, Hajicek, & Rubinstein, 2010), and speech materials (e.g., Miller, Heise, Lichten, 1951). The first study in this dissertation was directed toward the third area, the testing formats of speech materials. The second study expanded the findings of the first one, exploring a refined scoring method in analyzing listeners' word recognition (WR) test results.

In spoken word recognition tasks, listeners are presented with speech sounds and instructed to either repeat what they heard without cues other than the target speech sounds (open-ended or open-set testing) or to select their answer from a given list of word choices (multiple-choice or closed-set testing) presented in either pictorial or orthographic form. The first study quantified the precision of open- and closed-set WR tasks in identifying a change in hearing. Two different approaches were used to simulate a hearing loss – (1) sinewave-vocoded speech for simulating the degraded spectral resolution from hearing loss and (2) speech presented in noise for simulating the decreased audibility of speech sounds.

The second study examined a fine-grained method for analyzing WR performance to learn if this kind of analysis could reveal details about a person's hearing that are not present in the traditional analysis. This study was motivated by the findings

in the first study where listeners' performance on an open-set WR test scored by the percentage of correct words (or the word score) in the most adverse vocoded speech condition was near or at floor (0%) while their performance scored by the percentage of correct phonemes (or the phoneme score) was considerably higher than 0%. What was unknown before the experiment was whether this higher score was higher than what could be achieved by chance. Computer simulations were completed to answer this question.

Chapter 2: Diagnostic Precision of Automated Forced-Choice Word Recognition Tests

Hearing loss is a chronic health condition and the most common cause of disability that affects over 640 million people worldwide (World Health Organization, 2008). It is estimated to become one of the leading causes of the global burden of disease in 2030, primarily due to a growing global population with increasingly long-life expectancies (World Health Organization, 2008). In the United States alone, the annual financial cost of hearing loss is estimated to be between $154 billion to $186 billion, and yet 50% of the costs could be recovered when hearing loss is properly diagnosed and treated (Kochkin, 2007). However, the access to hearing healthcare service is limited due to a growing population of persons with hearing loss and a severe shortage of hearing healthcare professionals. This results in many underserved communities around the world in desperate need of diagnostic resources for hearing evaluation (Swanepoel et al., 2010). To address the dilemma, automation of hearing evaluation seems appealing.

In a typical hearing evaluation, the amount of hearing loss is quantified using a tonal threshold task which only accounts for the magnitude of the loss. Plomp (1986) proposed that effects of hearing loss could be quantified into an attenuation factor and a distortion factor. The attenuation factor of hearing loss represents decreased audibility of sounds and is clinically evaluated by the pure-tone threshold measure. On the other hand, the distortion factor refers to decreased clarity of sounds and can be directly assessed by speech-recognition tasks.

Evidence has supported Plomp's model (1986) of hearing loss, indicating that threshold sensitivity (e.g., pure-tone thresholds) and suprathreshold performance for

speech (e.g., word recognition) can be differentially affected by pathological conditions, and therefore both should be evaluated separately. For instance, studies reveal that individuals with Ménière's disease, auditory neuropathy, and acoustic tumor can have disproportionally poor word recognition scores for their amount of pure-tone threshold elevation (Chung, Hall, Buss, Grose, & Pillsbury, 2004; Starr, Picton, Sininger, Hood, & Berlin, 1996; van Dijk, Duijndam, & Graamans, 2000). By contrast, there are some pathological conditions where the word recognition score is better than predicted by pure-tone thresholds (e.g., a person with Norrie disease; Halpin & Sims, 2008).

Clinical assessment of speech understanding in the United States is time consuming, labor intensive, and its diagnostic accuracy is subject to human errors in scoring, although the errors for scorers with hearing within normal limits are small (Han, Schlauch, & Rao, 2014; Nelson & Chaiklin, 1970). In typical testing, an audiologist plays pre-recorded words to a listener at a fixed suprathreshold level and instructs the listener to repeat back each word. Responses are judged by the audiologist and a percentage correct word score is computed. This procedure is also known as the open-set response format of word recognition because the listener's response is not restricted to a limited list of choices. By contrast, for the closed-set format the listener is given a restricted list of words items to select among and only one is correct; the other choices are foils.

Due to time constraints, audiologists tend to present routinely only 25 words to evaluate word recognition performance (Martin & Sides, 1985). However, studies suggest that the number of words must be sufficient in order to adequately evaluate the listener's capacity and 25 words provide insufficient precision for making clinical

decisions (Carney & Schlauch, 2007; Schlauch & Carney, 2018). These challenges in clinical audiology with regard to test time, scoring accuracy, and precision make the automation of word recognition procedure appealing. An effective automated procedure will reduce human effort and improve diagnostic accuracy.

To date, only a few studies have assessed automating the clinical testing procedure for speech understanding. A study by Deprez, Yilmaz, Lievens, and Van Hamme (2013) used a computer speech recognizer (automatic speech recognizer or ASR). During a typical test in their protocol, which is common procedure in some European countries, the person under test repeats a sentence, and the audiologist records the number of keywords correctly repeated. To make the assessment objective and repeatable, this study evaluated if an ASR can take the place of an audiologist in recording the number of keywords identified during the process of scoring. The results show that ASR only achieved a keyword detection rate (the rate of correctly detecting correct responses) of only 88.8% with a false alarm rate (the rate of identifying incorrect responses as correct) of 11.2%. Even when speaker variability is accounted for, the keyword detection rate by ASR slightly increased to 90.7%. The false alarm rate remained approximately 10%. This is not an isolated finding because the current literature indicates that ASR still remains deficient in adapting to natural variation in speech, such as foreign accents (Elfeky, Bastani, Velez, Moreno, & Waters, 2016; Sahu, Dua, & Kumar, 2018).

Francart, Moonen, and Wouters (2009) attempted to automate the clinical speech task by having participants type their responses into a computer. Their protocol used an autocorrection algorithm to account for misspellings and the score was derived entirely

by computer based on the spell-corrected responses. In their automated computer-driven program, listeners were asked to listen to recorded monosyllabic words in Dutch via headphones, and then type what they heard using a keyboard. Typed responses underwent autocorrection to attempt to account for spelling errors, and the scores obtained were then compared to manual scoring by a human tester. They concluded that, while the autocorrection algorithm and human testers have comparable accuracy in scoring written responses, this method can be challenging to apply in other languages (e.g. English) where the correspondence between phonemes and graphemes is not as strict as Dutch. This method may also be limited in a clinical setting in general. Patients are expected to have sufficient level of literacy and to know how to use a computer keyboard to convey their responses.

A third approach to automating the clinical word-recognition procedure is a closed-set response task. A closed-set word recognition task restricts listeners' responses by providing a list of word options to identify their answer, such as California Consonant Test (Owens & Schubert, 1977), Word Intelligibility by Picture Identification (Ross & Lerman, 1970), and Modified Rhyme Test (MRT; House, Williams, Hecker, & Kryter, 1965). For instance, in the MRT (House et al., 1965), listeners are given a list of six alternatives that are rhymed and only differ by one phoneme for each test word. When a recorded word is presented to listeners, they are instructed to identify the target among the six alternatives, typically displayed orthographically on a computer monitor. The implementation of a closed-set task eliminates subjective judgments from clinicians in scoring and reduces the dependence of patients' speech production and their ability to spell. A possible workaround for participants with limited literacy skills is to use a

picture pointing task where each of the words is represented by an image, as in pediatric implementations of this closed-set approach. Therefore, use of a closed-set response format can be the potential solution to many of the aforementioned limitations in administering and scoring open-set word recognition tasks.

There are not any published studies, to our knowledge, that quantify the relative precision and efficiency of open-set and closed set tasks for identifying a change in hearing. The goals of the present study were (1) to quantify the diagnostic accuracy of open- and closed-set word recognition tasks for identifying a change in hearing and (2) to examine whether two different approaches for simulating a hearing loss – sinewave-vocoded speech and speech presented in speech-spectrum noise – yield similar results for relating WR performance among open- and closed-set tasks.

Method

Participants

Seventy normal-hearing, native English-speaking adults (54 female and 16 male) with audiometric thresholds of less than 20 dB HL at octave frequencies between 250 and 8000 Hz, participated in this study. Their ages ranged from 18 to 37 (median age 21). All experimental protocols were approved by the Institutional Review Board of the University of Minnesota. All participants were compensated $10 for a single 1-hr session and provided informed written consent prior to participation.

Stimuli

Words from the MRT (House et al., 1965) and the NU-6 (Tillman & Carhart, 1966) were used in this study.

MRT. Stimuli from the MRT (House et al., 1965) contain 50 ensembles with 6 monosyllabic words per ensemble, resulting in 300 words total. Words in each ensemble are rhymed and only differ by one phoneme. The 300 stimuli of the MRT used in the current study were digitized at a sampling rate of 48,000 Hz and downloaded from the database of Public Safety Communications Research ("Modified Rhyme Test Audio Library," 2015). The database contains nine versions of spoken stimuli made by different speakers (4 female and 5 male). A judgment on the degree of accent, naturalness, and clarity of speech across the recordings for the nine speakers was completed by the first author and two native English-speaking adults with experience directing speech recognition studies. Based on this review, the recordings produced by the third female English speaker (labeled as "F3" in the database of Public Safety Communications Research) were selected as stimuli for this experiment.

Each audio stimulus in the MRT word pool was then edited in a three-step procedure by the first author. First, the carrier phrase ("please select the word") was removed from each of 300 audio files using acoustic editing software (Audacity, version 2.0.5.0).

Second, due to the coarticulation between the target stimulus and the carrier phrase, each word file was then treated by adding a cosine-squared ramp (range: 10 – 230 milliseconds; mean: 46 milliseconds) at the onset of a word to remediate the unnaturalness caused by isolating a word from a sentence. In order to evaluate the naturalness and clarity of these modified audio stimuli, three native American-English speakers were recruited to identify each word in quiet and to comment on the quality of each word. Their recognition performance and feedback were then used to refine the

stimuli. The finalized version was tested in quiet to ensure the intelligibility of stimuli. A native American-English male adult speaker was recruited for the testing. Out of 274 unique words, he only missed one word ("teen" for "team"). Multiple additional listeners were tested on this word in quiet and all identified it correctly. This suggests that when no background noise or any spectral distortion is applied, the edited target words were highly intelligible.

Third, the root-mean-square (rms) level of the stimulus words was equated. The amplitudes of the original 300 digital files showed rms values within 1.5 dB when the carrier phrase and the word were analyzed. On the other hand, the range of rms levels was 8.9 dB for word-only files. After equating for rms amplitudes for the isolated words, minimal level differences remained for word-only files, the stimuli used in the experiment.

After equating for rms level, duplicate words were removed. An evaluation of the 300-word pool shows only 274 unique words in the MRT, including two pairs of homophones in the pool ("peel" and "peal"; "heel" and "heal"). After removing duplicates and one of two homophones, a subset of 272 words remained for later use in the MRT open-set tasks.

NU-6. Monosyllabic, consonant-nucleus-consonant (CNC) words from Northwestern University test #6 (NU-6) created for clinical assessment (Tillman & Carhart, 1966) also served as stimuli. NU-6 words were recorded materials (Q/Mass NU-6 lists 1D, 2D, 3A, 4A) that have been standardized and widely used clinically. Stimuli were spoken by a male talker with a carrier phrase "say the word…" in front of each stimulus word. The carrier phrase in the NU-6 was kept because this is a

standardized recording for clinical assessment. Four lists of fifty CNC words from the NU-6 were used in the current study and were transferred from a compact disk. On the Q/Mass recording, each list of fifty words was contained in a single audio file. The experimenter manually divided each recorded list into 50 audio files, resulting in 200 digitized files (4 lists × 50 words), one for each NU-6 word.

Simulated Hearing Loss

To simulate perceptions of decreased audibility and decreased spectral resolution, words were presented either in background noise or were processed with a speech vocoder. These methods for simulating hearing loss have been used in numerous other studies (e.g., Dorman, Loizou, Fitzke, & Tu, 1998; Lum & Braida, 2000). The power of this approach for our study is that listeners with hearing within normal limits can be assessed for a simulated change in hearing and the same change can be presented to all of the participants.

Noise. In the noise condition, each target word was presented in speech-shaped noise and bracketed by a 500-millisecond noise before and after the word segment. The spectrum for the speech-shaped noise was created separately for each speech material (i.e., the NU-6 and the MRT). Broadband noise was filtered to have the long-term average speech spectrum for the type of target words. The level of the target speech was set to 65 dB sound pressure level (SPL) to closely match the level of normal, conversational speech. The noise level was adjusted in a range of 57 to 69 dB SPL to represent five signal-to-noise ratios (SNR) (-8, -4, 0, and 4 dB).

Sinewave vocoder. To simulate decreased spectral resolution, each target word was processed through a sine-carrier vocoder, which simulates some aspects of

perception experienced by cochlear implant users. The vocoder was implemented in MATLAB® (The MathWorks, Inc.). The signal was first split into 2, 4, 6, or 8 logarithmically spaced frequency bands (or channels) having center frequencies between 300 to 6000 Hz at two extremes (Table 1). The intensity envelopes were then used to modulate a pure tone at the center frequency of the respective channel. Finally, the modulated pure tones were summed and scaled to produce a presentation level 65 dB SPL to each ear.

Table 1. Center frequencies (Hz) for four vocoder conditions

	Center Frequencies							
Number of channel	1	2	3	4	5	6	7	8
2	300	6000						
4	300	814	2210	6000				
6	300	546	994	1810	3296	6000		
8	300	460	706	1083	1662	2549	3911	6000

Procedure

The MRT words were the primary stimuli in this study. The MRT words are normally presented in a closed-set format but in this study open- and closed-set MRT conditions were collected. Because of the primary interest in clinical applications, data from open-set NU-6 words were obtained to provide context.

There were eight participant groups in this study. Table 2 summarizes the number of participants in each group. Participant groups differed by the speech material

(MRT or NU-6), listening condition (noise or vocoder), and response format (open-set task, 4-alternative closed-set task, or 6-alternative closed-set task). Each listener was randomly assigned to one of eight participant groups. Initially, five participants completed each of the eight listening conditions. The initial goal was to define average psychometric functions to be used in a simulation. After this initial data collection was completed, additional participants were recruited for the 6-alternative closed-set tasks and the open-set tasks for the MRT words to improve the precision of simulations and individual participant analyses for these important conditions.

Participants started with ten practice trials before beginning the experiment. Word trials in both the practice and experimental tasks shared the identical response format and type of simulated hearing loss. However, words in the practice and experiment were retrieved from different speech materials. If stimuli from the MRT were used in the experimental task, words from the NU-6 were used for practice trials, and vice versa. Although there were 24 words overlapping in both speech materials, test words in the practice trials did not repeat in the experimental task. All stimuli were presented binaurally via headphones (Sennheiser, HD580) to the listener seated in a double-walled, sound-isolated booth. The tasks were programmed in MATLAB® (The MathWorks, Inc.). Listeners were instructed to follow the prompts on the screen to proceed throughout the practice and experimental tasks.

Participants in each group undertook four different levels/amounts of simulated hearing loss throughout the experiment in a blocked form where the level/amount was fixed within a block. To simulate decreased audibility, participants listened in four levels

of background noise (SNRs of -8, -4, 0, and 4 dB). Four levels of spectral distortion were simulated: 2, 4, 6, and 8 channels. The order of the four blocks was randomized.

Table 2. Number of participants in each group

		Response format		
Speech Material	Simulated condition	4-altermative closed-set task	6-altermative closed-set task	Open-set task
MRT	Vocoder	5	10	10
	SNR	5	15	15
NU-6	Vocoder	-	-	5
	SNR	-	-	5

Open-set tasks. Listeners in the open-set tasks were asked to repeat each word after it was presented. Both their verbal responses and facial image were recorded using a webcam (Logitech, C920 HD Pro) and used for later verification of typed responses. Participants were instructed as follows:

> In this experiment, you will be presented with distorted spoken words or words in background noise (*the appropriate instruction was given based on the listening condition*). After a word is presented, your task is to (1) face the camera and say the word aloud, (2) type your answer in the response box on the computer screen, and (3) hit "Enter" on the keyboard to proceed. If you are not sure of the word that is said, or if you hear just part of the word, make your best guess.

In the groups using the MRT words, participants were presented with 272 words and prompted to take a break after every 68 trials, resulting in four blocks for each

experimental task (4 blocks × 68 words = 272 words). Each block represented one of four levels of simulated cochlear hearing loss.

In the groups presented with the NU-6 words, participants were presented with a total of 200 words and prompted to take a break for every 50 trials, resulting in four experimental blocks where 50 words in each block represented an entire list of NU-6 words. Each of four blocks in each NU-6 task represented a degree in simulated hearing loss.

Closed-set tasks. There were four participant groups receiving word recognition tasks in a closed-set format. Participants were instructed as follows:

> In this experiment, you will be presented 300 spoken words that are distorted or presented in background noise (*the appropriate instruction was given based on the listening condition*). After a word is presented, your task is to identify the word that is being said using the mouse to select your answer from a list of choices displayed on the computer screen. If you are not sure of the word that was said, or if you hear just part of the word, make your best guess.

In the closed-set tasks, participants were presented with 300 words from the MRT and prompted to take a break after every 75 trials, resulting in four blocks for each experimental task (4 blocks × 75 words = 300 words). In a 6-alternative task, each test word (e.g., “went”) was presented through the headphones along with a visual presentation of six rhymed words (e.g., “went”, “sent”, “bent”, “dent”, “tent”, “rent”) displayed on the computer screen. For a 4-alternative task, only four rhymed words (e.g., “went”, “bent”, “dent”, “tent”) were available for selection for each test word (e.g., “went”). For each test trial, word alternatives were visually presented, followed by the

audio presentation of a target word; the visual presentation of alternatives remained on the screen throughout each trial but was greyed-out initially so that participants were unable to make a choice until the audio presentation of a target word was presented.

Scoring

Performance in open-set tasks. Participants' typed responses were recorded into a spreadsheet by the MATLAB® (The MathWorks, Inc.) program and scored offline by the main author and two trained research assistants. Audio recordings were accessed for ambiguous responses. Participant's responses for a given listening condition were scored as the percentage of correct words and percentage of correct phonemes. Scoring reliability was monitored by having a second grader crosscheck the scores of a subset of participants (63% of the data), all discrepancies were resolved after discussion among graders.

Scored by word. When the performance was scored by word, a response was judged correct when the typed response matched the target words orthographically. In most cases, if any part of the participant's response to a word was incorrect, the entire word was scored as incorrect. However, two cases of disagreement between an orthographic response and a target word were scored correct – homophones and misspellings. A homophone was considered as a correct response. Responses were considered as typographical errors when participants' verbal responses match with the target word phonetically. There were 200 test words in the tasks using the NU-6, resulting in 50 words per listening condition, making each word contribute 2% to the score. There were 272 test words in the MRT tasks, resulting in 68 test trials per listening condition, making each word contribute 1.47% to the score.

Scored by phoneme. Rules in Appendix A were used to score the percentage of correct phonemes for the open-set word recognition performance. These rules were adapted from Schlauch, Anderson, & Micheyl (2014) to fit the current tasks. The two speech materials (i.e., the NU-6 and the MRT) in this study differ in the number of phonemes in each test word. Each word in the NU-6 had three possible phonemes; therefore, the percentage of correct phonemes in each block was based on a total of 150 scored items (50 words × 3 phonemes). On the other hand, the number of phonemes in each MRT word ranged from 2 to 4 (mean: 3.05 phonemes), each block was based on a total of 208 scored items (68 words × 3.05 phonemes).

Performance in closed-set tasks. In each closed-set task, participants' responses were recorded automatically into a spreadsheet and the accuracy of each trial was scored by the MATLAB® (The MathWorks, Inc.) program. Responses were tallied to derive the percentage of correct words for each condition.

Results

Psychometric Functions

Figure 1 illustrates group-mean performance for the eight listening tasks for all 70 participants, distributed among conditions as shown in Table 2. Data for each condition were fitted using Probit analysis (Finney, 1952), which took into account binomial variability and the expected lower asymptote representing chance success for forced-choice conditions. The top panel in Figure 1 (A and B) presents the psychometric functions for open-set conditions for the two types of stimulus words (MRT and NU-6) for both types of simulated loss (vocoder and noise). Both types of stimulus words show monotonically increasing functions with an increased number of channels (left panel)

and a higher SNR (right panel). As has been reported by others (e.g., J. T. Gelfand, Christie, & Gelfand, 2014), percentage of correct phoneme scores are higher than those for percentage of correct words for a given stimulus condition. Further, performance for the NU-6 words was higher than that for the MRT words for the same listening condition.

The lower panels in Figure 1 (C and D) show the relationship among the closed- and open-set tasks for the MRT words. The data for the open-set MRT words in panels A and B are replotted in these panels. Both types of simulated hearing loss produced similar data. The closed-set conditions produced much shallower slopes than those for the open-set conditions. Further, performance was higher for the closed-set than the open-set tasks for the same stimulus condition. The slope of the 6-alternative closed-set condition in this study is nearly identical to the one reported by Letowski and Scharine (2017) when compared for a similar range of SNRs. Further, the shallower slope in noise and vocoderized speech for the closed-set tasks than for open-set monosyllabic word tasks is consistent with the findings of Williams and Hecker (1968).

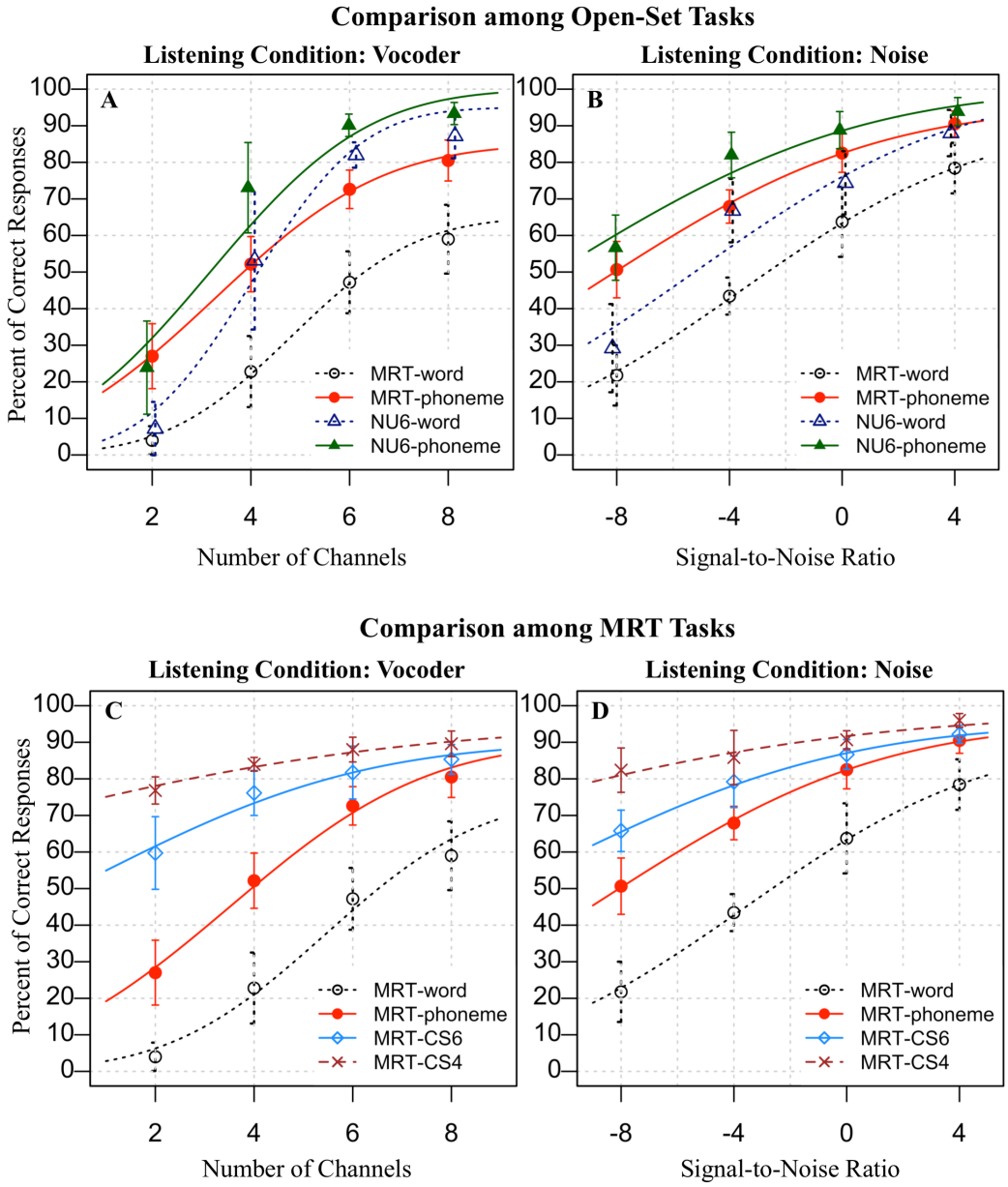

Figure 1. Psychometric functions for open-set tasks and for MRT tasks in two degraded listening conditions

Computer simulation: True positive rate in detecting a change in hearing

Studies have revealed that the variability of speech-recognition scores for monosyllabic words can be modeled as a binomially distributed quantity (Thornton & Raffin, 1978). Tables of 95% confidence intervals have been derived based on these statistics for clinical application to assess whether two consecutive speech scores are significantly different for word scores (Thornton and Raffin, 1978; Carney & Schlauch, 2007) and for words scored by the percentage of correct phonemes (Schlauch and Carney, 2018).

If one assumes that the group-mean psychometric functions in Figure 1 represent the underlying "*true*" scores for an average listener, these scores for different stimulus conditions can be used in a simulation to compare the relative precision of the closed-set and open-set tasks for identifying a change in hearing. For example, as the SNR is increased from -8 to -4 dB, the interpolated performance for the MRT open-set task increased nearly 19%, from 3.98% to 22.8%, for the word score. By contrast, for the same stimulus conditions, performance for the 4-alternative closed-set task only increased 7%, from 76.82 to 83.98%. To learn about the relative precision of these tasks for identifying that 4-dB improvement in SNR, it is necessary to consider the variability of estimates of scores that might be observed for someone with actual scores corresponding to the true performance.

To conduct the simulation, we simulated the range of expected scores for single lists of 25, 50 and 125 test items. One hundred twenty-five items were simulated because the variability associated with a 50-word list of CNC words analyzed by the percentage of correct phonemes is equal to 125 and not 150 items due to a lack of

independence of the phonemes[1] (Boothroyd & Nittrouer, 1988; Gelfand et al., 2014; Schlauch & Carney, 2018). Fifteen thousand values were drawn from a binomial distribution that had parameters corresponding to the list length and *p*, the proportion of correct responses obtained from the psychometric function. The simulation was completed for multiple values of *p* representing interpolated values of scores on the psychometric function.

Figure 2 illustrates distributions of discrete scores obtained by simulation for a 4-dB increase in SNR for open-set tasks scored by word and phoneme and for 4- and 6-alternative closed-set tasks. The left-most distribution in each panel represents the range of expected scores for the -8 dB SNR condition. The right-most distribution represents the range of scores for a -4 dB SNR. To assess the ability of a task to identify an improvement in SNR, a pass-fail criterion was established that represents the 95th percentile for the 15,000 simulated scores for the lower distribution (i.e., the distribution for the -8 dB SNR condition). Any score exceeding that value (the 95th percentile) was judged to be an improvement in performance. This pass-fail criterion results in a 5% false-positive rate. The true-positive rate depends on the overlap in the distributions. For this example, the simulation based on 50-word lists in the open-set tasks (Figure 2A and B) yielded much higher true-positive rates than the closed-set tasks (Figure 2C and D).

Simulations, as described above, were completed for a range of vocoder channels and SNRs for each of the tasks. Figure 3 illustrates true-positive rates (assuming a fixed 5% false-positive rate) for a variety of conditions for each task. Two major trends were revealed by the simulation analyses in Figure 3. First, the number of test items contributing to a score determines the diagnostic precision of a WR task. The true

positive rate for most WR tasks improved as the number of test items increased from 25 to 50 and from 50 to 125, when not limited by ceiling effects. For instance, when the scores in the 6-alternative closed-set tasks were obtained from testing only 25 words, there were only 38% of scores in the SNR of -4 dB were distinguishable from the baseline performance of -8 dB SNR (Figure 3D). However, the true-positive rate increased to 95% for the same task when the number of test items increased to 125 (Figure 3F).

Second, with the same number of test words, the closed-set tasks are relatively poor in identifying the smallest change in hearing (i.e., 2 to 4 channels in the vocoder condition and -8 to -4 dB change in the noise condition) as compared to the open-set tasks. For instance, with 50 test words, the open-set tasks correctly identified 93% and 99% of cases with a change in hearing from -8 to -4 dB SNR when analyzing the performance by the number of correct words and phonemes, respectively (Figure 3E). On the other hand, only 15% and 65% of cases of a hearing change were identified by the 4- and 6-alternative closed-set tasks, respectively (Figure 3E).

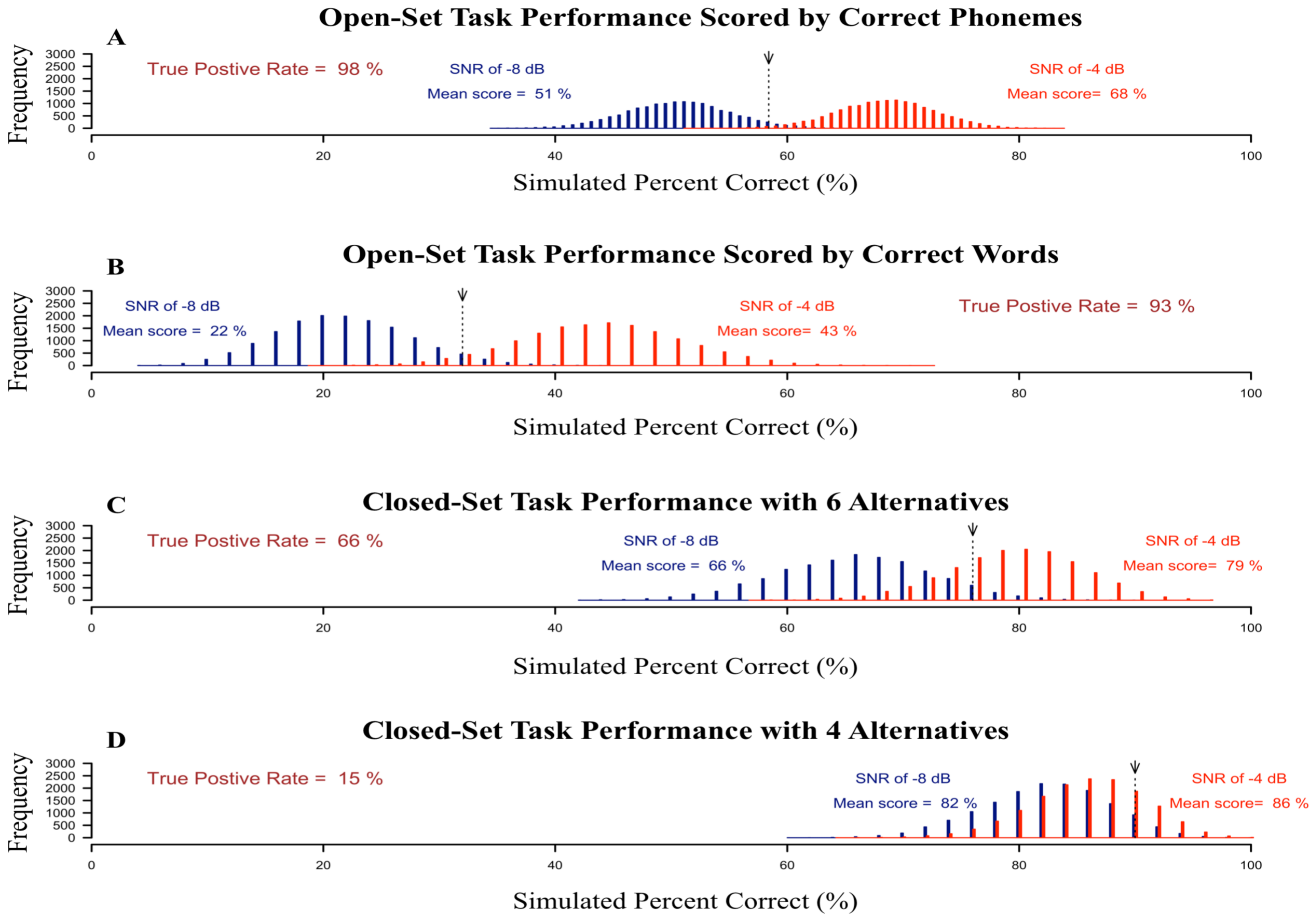

Figure 2. Binomial distributions for representing expected speech recognition scores on 50-word tasks at SNR of -8 dB and -4 dB

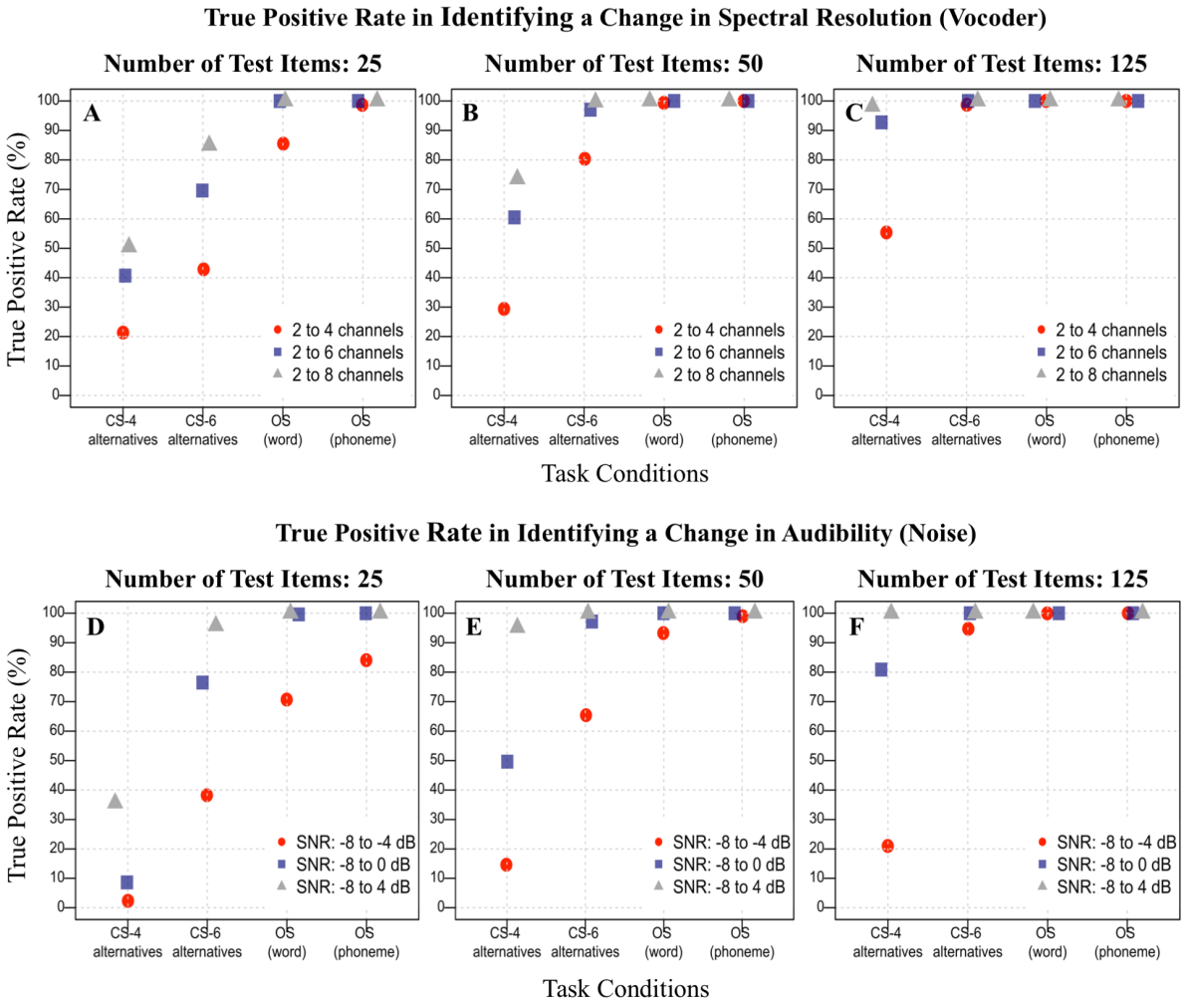

Figure 3. Computed true positive rates based on 15,000 simulated observations at each listening level using the mean values from the fitted functions for 25-, 50-, 125-word tasks

A test using individual data

To test the validity of the conclusions drawn regarding the relative precision of the tasks that were derived with simulations based on group mean data, we evaluated individual data for the open-set and the 6-alternative closed-set task in noise. Each had 15 participants. True-positive rates for identifying a change in hearing were obtained by comparing scores for an individual at two SNRs and assessing whether the score for the higher SNR exceeded the 95% confidence interval for the score obtained at the lower SNR. For each of the 15 participants, two pairs of scores based on 25-word lists were assessed at different SNRs which resulted in 30 comparisons for each SNR difference. Analyzing two, 25-word lists for each listening condition was possible because each point on the psychometric functions of the MRT open-set and the 6-alternative closed-set tasks were obtained using 68 and 75 words, respectively. For each condition, the first 18 test trials were removed, and the following 50 trials were selected and divided into two 25-word lists. Published 95% confidence interval tables (Carney and Schlauch, 2007) and the confidence intervals tabled in Appendix B[2] were consulted to determine if the score for the higher SNR was significantly different from a score at the lower SNR. The percentage of cases identified as different, out of 30 pairs, determined the true positive rate for detecting a change in hearing. Figure 4 reveals similar patterns of changes of true positive rates across closed- and open-set tasks for the simulated data obtained from 15,000 samples and the individual data from 30 observations, a finding that supports the validity of the computer simulation.

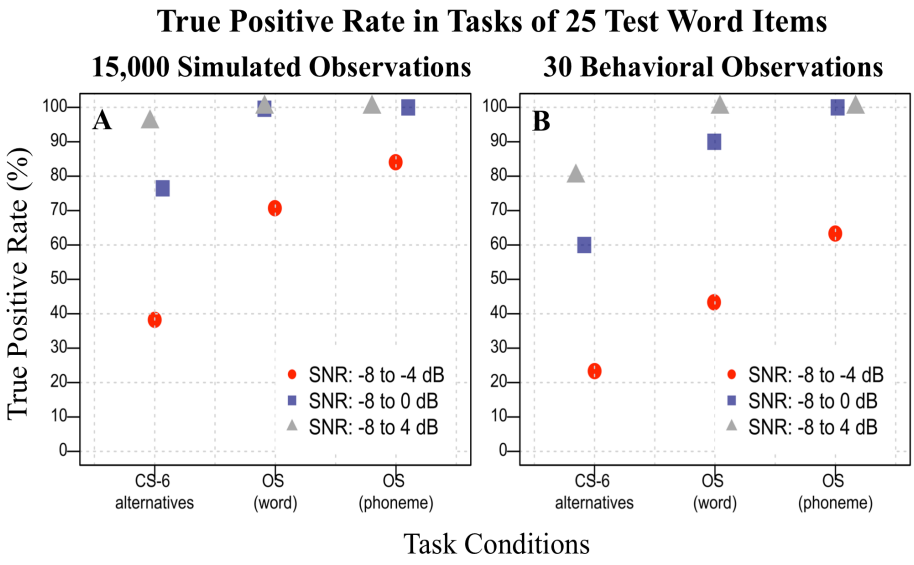

Figure 4. True positive rates for simulated data derived from 15,000 observations using the group-mean psychometric functions are compared with true positive rates obtained from individual behavioral data obtained using 95% critical difference tables for 30 pairs of scores obtained at 2 SNRs for 25 test words

A second computer simulation: Number of test trials for equivalent true positive rates

An additional simulation was conducted to provide a context for directly comparing the precision of the different tasks. It is well known that more test items leads to better test precision (Thornton & Raffin, 1978). The goal of this simulation was to establish the number of test items required to achieve identical precision for each of the tasks. The simulation data were run for a condition representing an increase in the SNR ratio from -8 to -4 dB for true-positive rates ranging from 50% to 98%. The data in Table 3 reveal that more than an order of magnitude increased in test items are needed for the 4-alternative closed-set task to achieve the same precision as the open-set task analyzed by phoneme.

Table 3. Number of test words necessary for equivalent true positive rates for each task condition with the false positive rate fixed at 5%

True positive rate	Number of test trials by tasks			
	Closed-set task with 4 alternatives	Closed-set task with 6 alternatives	Open-set task (scored by word)	Open-set task (scored by phoneme)
98%	1600	160	70	50
95%	1200	120	54	35
75%	620	64	26	18
55%	380	49	18	14
50%	329	32	15	12

Discussion

Two very different distortions of hearing loss yielded nearly identical results regarding the precision of open- versus closed-set WR tasks. The general result was based on computer simulation using group-mean average psychometric functions. A subset of the data for individuals in the speech-noise condition supported the results of the simulation derived from group-mean data. The relative precision of these tasks from worse to best is 4-alternative closed-set tasks, 6-alternative closed-set tasks, open-set tasks analyzed by percentage of correct words and open-set tasks analyzed by percentage of correct phonemes.

The finding that the closed-set tasks are less precise than those in the open-set tasks has been reported previously. Clopper, Pisoni, and Tierney (2006) found that closed-set tasks may be less sensitive than open-set tasks when the number of choices is low and the foils are not selected from "*hard*" words, which are ones that have a low frequency of occurrence and are phonetically similar to many high-frequency words. They argue that the task demands are simpler in a closed-set task because the lexical competition is based on the words selected as foils rather than the entire lexicon. Factors that contribute to the process of speech recognition include lexical competition and talker variability. Words with few lexical neighbors were recognized better and were processed more quickly in lexical decision and naming tasks than words with many neighbors (Luce & Pisoni, 1998). Meanwhile, speech produced by a familiar talker was processed better and recognized more quickly than speech by an unfamiliar talker (Murphy, 2002). Studies show a lack of effects of these factors on conventional closed-set word recognition tasks[3], implying that closed- and open-set word recognition tasks

are not evaluating the same construct of speech understanding (Clopper et al., 2006; Sommers, Kirk, & Pisoni, 1997). In a 4- or 6-alternative closed-set task using rhymed foils, the task demand for a listener is similar to a discrimination task where listeners compare the similarity among the target and the foil phonemes. In contrast, an open-set task requires listeners to recognize sufficient speech sounds in a word in order to correctly identify the target.

An analysis suggested by Green (1990) of psychometric functions can be applied to performance in open and closed set tasks to help explain the relative precision of these tasks in the present study. Green's analysis was intended to provide insights into adaptive threshold procedures for measuring detection and discrimination thresholds in psychoacoustic tasks. His goal was to identify a psychometric function "sweetpoint", the location on the psychometric function that minimizes variability for finding thresholds. Accordingly, the variability of a probability estimate is summarized by the binomial variance for a given probability divided by the slope of the psychometric function squared. Applied to the data in Figure 1, the closed-set psychometric functions have a much shallower slope than those for the open-set psychometric functions. As the slope becomes shallower, the percentage change in performance with a change in stimulus attenuation/distortion diminishes. It is this shallower slope that is primarily responsible for the poorer precision in the closed-set tasks. As suggested in the literature, the cause for the shallow slope in closed-set tasks is the level of chance success (Schlauch & Rose, 1990) and the cognitive demands of the foils (Clopper et al., 2006).

Based on our results, the 4-alternative closed-set task with rhyming words has limited clinical application because of its poor precision. The 6-alternative closed-set

task had much improved precision over the 4-alternative closed-set task and could still find useful applications. The MRT, which is a 6-alternative closed-set task, is used to assess the intelligibility of speech over communication systems (ANSI S3.2, 2009). The MRT does not require a scorer which makes test administration more economical and the scoring more accurate. The MRT also has a minimal practice effect from multiple exposures and needs nominal training from listeners (House et al., 1965), unlike open-set monosyllabic word tests, which when used in testing communication systems require a crew of listeners and scorers trained on the word lists. Naïve listeners can be used in studies to assess communication systems or clinical speech understanding using open-set WR tasks, but the small number of different words available for testing can place the limit on the number of listening conditions that can be examined (e.g, there are only 200 words in the NU-6).

In this study, the open-set test was found the most efficient test for identifying changes in hearing. The open-set test can be analyzed by the percentage of correct phonemes, which increases the test precision (Schlauch et al., 2014). Because the MRT foils only differ by one phoneme, an analysis by percentage correct phonemes is not an option to increase precision over analysis by word. That stated, the California Consonant Test (Owens & Schubert, 1977), which is a closed-set test, can be analyzed by distinctive features to increase test precision (Feeney, 1990).

This study suggests that open-set tests are more precise than closed-set ones in diagnostic evaluation; however, in reality an open-set test is not an option for all clinical populations. For instance, children with congenital hearing loss may be unable to produce accurately the speech sounds being assessed by the hearing test. The same is

true for non-native speakers of the language being tested as well as persons with speech disorders. Evidence suggests that a 12-alternative closed-set task with carefully selected foils is sensitive to differences in talker and lexical competition (Clopper et al., 2006) but the memory load required to process eleven foils may be too much for some clinical populations.

Conclusion

Consistent with the findings of Clopper et al. (2006), this study reports evidence that open-set tasks are much more sensitive than closed-set tasks (based on 4 and 6 choices) for identifying a change in hearing. For the same number of test words, scoring the open-set performance by the percentage of correct phonemes was found to be the most precise test. By contrast, the 4-alternative WR task was the least efficient in detecting a change in hearing. The goal of automating speech tests for clinical audiology is a laudable one and the use of closed-set type of tests is essential for some clinical applications, but given our current state of technology, this study suggests that open-set WR testing analyzed by phoneme provides a powerful tool for assessing the distortion component of sensorineural hearing loss.

Chapter 3: Developing Finer-Grained Methods for Analyzing Word Recognition Performance by Individuals with Cochlear Implants

Speech recognition tests have always played an important role in assessing perception with electrical hearing for listeners with cochlear implants (CIs). In typical cases of sensorineural hearing loss, there is a strong correlation, with some notable exceptions, between the amount of hearing loss (measured by the pure-tone audiometry) and the ability to understand speech. This makes the pure-tone thresholds a potential predictor in speech recognition performance. This method works so well with persons with mild and moderate losses (e.g., Pavlovic, 1984) that formal computational rules for predicting speech understanding have been incorporated into an American National Standard (ANSI S3.5, 1997). For individuals with CIs, however, the speech recognition tests become irreplaceable because individual variability in performance is large and the correlation is weak, at best, with non-speech measures of electrical hearing.

In order to evaluate CI listeners' longitudinal progress in speech perception, studies have been completed to explore the use of spoken word and sentence recognition materials for this purpose. For instance, since the Food and Drug Administration (FDA) approved the CI use in 1984, the open-set Consonant-Nucleus-Consonant (CNC) word test (Peterson & Lehiste, 1962) have been routinely used in diagnostic evaluations and for measuring the CI outcomes. However, the performance on the CNC word test was much poorer, on average, with early CI technology yielding the CNC scores at floor performance. As a consequence, open-set sentence recognition materials, such as the Hearing in Noise Test (HINT; Nilsson, Soli, & Sullivan, 1994), were developed that

were “easier” than the CNC word test and, as such, were able to document meaningful differences in performance in this population.

Thanks to remarkable advances in implant technology and speech processing strategies, CI listeners’ ability to understand speech has been largely improved since the formal introduction of CI in 1984. As a consequence, current CI technology makes the understanding of single words more possible while making the sentence task prone to the ceiling effects. Studies have shown that the HINT (Nilsson et al., 1994), a sentence task, was previously often used in determining adult CI candidacy and assessing post-implantation performance, was found subject to ceiling effects when presenting the sentences in quiet (e.g., Gifford, Shallop, & Peterson, 2008). To address this limitation, the AzBio sentences were then developed and recorded by multiple talkers at a conversational speech rate in contrast to the single talker in the “clear speech” mode in the HINT sentences (Spahr & Dorman, 2004). As a result, the AzBio sentences yielded lower scores than the HINT; however, on average, scores on the AzBio were still higher than the scores on the CNC word test in a large group of CI listeners and hearing aids users (Gifford et al., 2008). A recent study by Sladen et al. (2017) compared the use of the AzBio and the HINT sentences with the CNC words for determining the CI candidacy and long-term speech perception outcomes. Two major findings were concluded in Sladen et al. (2017). First, the results showed an overall trend for candidacy to be based on monosyllabic word recognition scores (e.g., the CNC word test). Second, 60% of participants’ performance on sentences reached ceiling values after only 3 months of implant use while none of the participants reached a score of 80% on the CNC word test even 12 months after the surgical operation. These findings

suggest that the monosyllabic words are more appropriate than the sentence materials for measuring the long-term postoperative speech recognition performance.

Due to the idiosyncratic nature of CI listeners, some listeners may yield high scores on the CNC words with today's improved technology while others may still produce very low scores that represent floor performance. If we assume a condition where a person does not hear anything that person's chance in getting any CNC words correct is effectively zero. The same outcome is not true for words scored by the percentage of correct phonemes. If that person with no hearing ability were to respond with a CNC word following each presentation of a target word, the score would likely be higher than zero because some of the phonemes in the responses would line up with those in the target words. To our knowledge, there are no studies to date addressing chance performance for the percentage of correct phonemes for monosyllabic words.

Studies have shown that scoring WR performance by the percentage of correct phonemes provides advantages over scoring by the percentage of correct words. One advantage is the increased precision in diagnostic evaluations for roughly the same time in test administration (Schlauch, Anderson, & Micheyl, 2014; Schlauch & Carney, 2018). Meanwhile, performance on the monosyllabic words has not been found suffering from ceiling effects over time under a variety of implantation options, such as bilateral, unilateral, or bimodal implantation (Gifford et al., 2008; Sladen et al., 2017), showing the value of a WR test as the tool for longitudinal evaluations of speech performance. A WR score also could be at or near zero percent correct even if a person is able to recognize a significant percentage of phonemes correctly. If this is true, analyzing scores

by the percentage of correct phonemes would extend the range of clients that could be assessed using monosyllabic words as stimuli.

The purpose of this study is to extend the utility of WR tasks scored by the percentage of correct phonemes. A series of simulations were conducted to determine the range of chance performance on this measure for a clinical, monosyllabic WR test for adult listeners, the Northwestern Auditory Test No.6 (Tillman & Carhart 1966; NU-6), which was derived from the CNC word test (Peterson & Lehiste, 1962). There were three research goals in this study as follows.

The first goal was to explore the effect of NU-6 list number on the chance performance scored by percentage of correct phonemes (phoneme chance performance). The four, 50-word lists that constitute the NU-6 test were phonetically balanced (Tillman & Carhart 1966), meaning that all phonemes appear in a list with a frequency that approximates the frequency with which they are used in the English language. Since the four lists were equivalent, it is hypothesized that similar distributions of phoneme chance performance should be observed among the NU-6 lists.

The second goal was to examine the effect of word frequency on chance performance for the percentage of correct phonemes. Words in the NU-6 occurred at least once per million words in printed English, meaning that words that were relatively infrequent were not included in the development of the NU-6 (Tillman & Carhart 1966). With a lexicon of only frequent words available as the pool of words used to "guess" the target word, listeners might have better chance in guessing phonemes in the NU-6 than listeners with a lexicon of frequent and infrequent words available for WR. Therefore, it

was hypothesized an effect of word frequency should be observed on phoneme chance performance.

The third goal was to evaluate the effect of phoneme position on chance performance. Words in the NU-6 were represent a subset of the CNC test words (Peterson & Lehiste, 1962) so they are also in the CNC form. Since there are fewer vowels than consonants in English and the NU-6 lists (15 vowels and 22 consonants), phoneme chance performance was predicted to differ by the phoneme category with relatively higher guess rate for vowels than for consonants. A related goal is to extend the clinical application of the chance performance by phoneme categories. Researchers have been studying differences in the perception of vowels and consonants among CI listeners. For instance, Munson, Donaldson, Allen, Collison, and Nelson (2003) found that CI listeners varying in overall performance tended to differ quantitatively and less qualitatively in their phoneme misperception: listeners struggled with the place feature more than other consonant features (voicing, manner, and duration). They also were more challenged in perceiving the height and r-coloring features than other vowel features (tenseness, pitch, and backness). To improve the clinical utility of WR scores, an objective for this study was to evaluate the performance differences between two phoneme categories (vowels and consonants) and three phoneme positions (beginning, middle, and final) at the level of individual listeners in a simulated CI condition.

Method

Word corpuses

Target words. Four lists of 50 monosyllabic words (a total of 200 words) from the NU-6 (Tillman & Carhart, 1966) were used as the pool of target words. All words

were in the CNC form. Each NU-6 word was transcribed using the Speech Assessment Methods Phonetic Alphabet (SAMPA) codes and was associated with a word frequency index using the Hyperspace Analogue to Language (HAL) frequency norms (Lund & Burgess, 1996).

Response words. A corpus of CNC words were retrieved from the English Lexicon Project (ELP) online database (Balota et al., 2007). A step-by-step procedure (Appendix D) was followed to create the list of 2269 CNC words for this study. Each word in the ELP corpus contains 1) the transcription codes based on the SAMPA and 2) the log of word frequency reported by the HAL frequency norms (Lund & Burgess, 1996). Both the transcription code and the word frequency index were necessary in later simulation of phoneme chance performance. The finalized ELP corpus was treated as the word pool from which random responses to target words were retrieved.

Computer simulation

The simulation was performed using *R* (Ihaka & Gentleman, 1996) and was done in the following way (Appendix E):

1. Determine the list used for target words by choosing the word list(s) among four from the NU-6 (List 1, 2, 3, and/or 4) and the size of the test list by selecting the list length (e.g., 50 words for a full list, and 100 words for 2 full lists).
2. Determine the pool of response words (i.e., the ELP corpus) by factoring in the parameter of word frequency using the HAL frequency norms (Lund & Burgess, 1996). Two pool sizes were used. One pool of words included all of the CNC words in English; the other included only frequently occurring words that were

higher than that of the average of word frequency for the NU6 words (Log HAL of 9.68).

3. Randomly retrieve, without replacement, words from the finalized ELP corpus as response words. The number of random response words matched the number of target words as determine in Step 1.
4. Pair each random response word (from Step 3) with a target word from the NU-6 (from Step 1) accordingly and compute the number of phonemes correctly detected for each word pair.
    a. Example: If the target word is '*fall*' and the response word is '*face*', the phoneme score is 1 out of 3 for this trial.
    b. Example: If the target word is '*raise*' and the response word is '*leaf*', the phoneme score is 0 out of 3.
5. Compute the percent phonemes correct for each test list.
6. Repeat the aforementioned Step 3 to 5 for 15,000 times to generate a large number of percent scores of correct phonemes for random responses.
7. Identify the cut-off score on the derived distribution by searching the score at the three standard deviations above the mean and set it as the criterion for the significance test.

Comparison with human performance data for NU-6 words processed using a 2-channel vocoder condition

To demonstrate the clinical application of simulated findings in the individual performance on WR performance, participants' performance on the NU-6 were obtained from Yu & Schlauch (2018). Five participants with normal hearing were instructed to

listen to the NU-6 words from the recording in a simulated CI listening condition where the signal was spectrally divided into two, four, six, or eight broad spectral channels and processed through a sinewave carrier vocoder. Individual performance in the most adverse listening condition (2-channel vocoder) were revisited and analyzed by the percentage of correct words, phonemes, and phonemes by position.

Simulation Results

Effect of NU-6 list

One simulation condition examined if the chance percentage of phonemes correct for the NU-6 WR task was dependent on a chosen test list from the NU-6. The simulation showed nearly identical shape of distribution and statistics for scoring by phonemes for all four lists. This suggests the phoneme chance success on the NU-6 task is not affected by any given lists. For all four lists, chance performance ranged from 0 to 17%; Figure 5A shows the example using List 1 as the target word list. A pass-fail criterion was selected as a score was three standard deviations above the mean. Scores better than the criterion suggest that the performance is unlikely due to chance.

Since all four lists yielded the same result, the following analyses were conducted based on List 1 of the NU-6 and are assumed to be representative for the other lists.

Effect of word frequency

To evaluate if the chance success rate was affected when only words with frequent occurrence were available for responses, simulations were conducted by factoring the word frequency parameter into the selection of response words from the ELP corpus. The word frequency parameter was set based on the average frequency of

occurrence for words in upper 50th percentile of the NU-6 words. The averaged log of word frequency for the NU-6 was 9.68 on the HAL frequency norms (Lund & Burgess, 1996). Using this averaged word frequency as the parameter, about 78% of the words in the ELP corpus were considered relatively infrequent words as compared with the NU-6. The removal of these infrequent words from the ELP corpus resulted in a much smaller response pool of 492 words. Figure 5B illustrates the distribution of phoneme chance performance when low-frequency words were eliminated from the response word pool. A comparison of the distribution in Figure 5A reveals that the distributions have a similar range, mean, and modal of phoneme chance scores on a WR task for both conditions.

When inspecting the distribution of scores, the chance success for getting one word correct out of fifty words were nearly doubled from 3% to 6% of the 15,000 simulations after reducing the pool of response words from 2269 to 492. The chance of getting two words correct remained at 1 in 1000 samples or less for the original and reduced pool of response words. This suggests that even though the chance success for getting words correct was slightly higher with frequent words (3 versus 6%), the averaged phoneme chance performance on a WR task is not predicted to be affected by how frequent or infrequent words are available in a listener's lexicon.

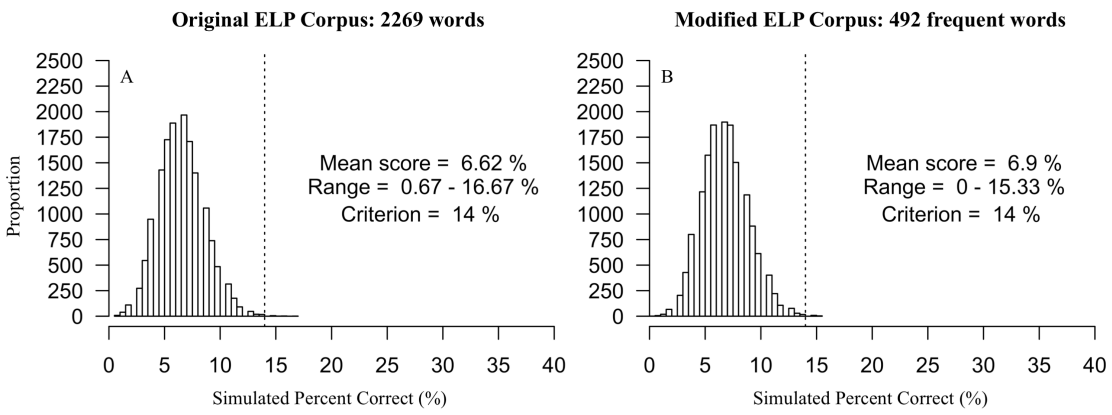

Figure 5. Effect of word frequency on phoneme chance performance

Effect of phoneme position

To explore the effect of phoneme position on the chance performance, the percentage of correct phonemes identified in the initial, medial, and final position for each list were computed. Figure 6 shows that chance performance was affected by the phoneme position and the phoneme category. The mean score was the highest for the medial vowel, followed by the final consonant and the initial consonant. The criterion also differed by the phoneme position accordingly. These findings and the following example demonstrate a refined way to analyzing the WR performance by phoneme position and provide a potential way to evaluating the strength and weakness of listeners' speech perception.

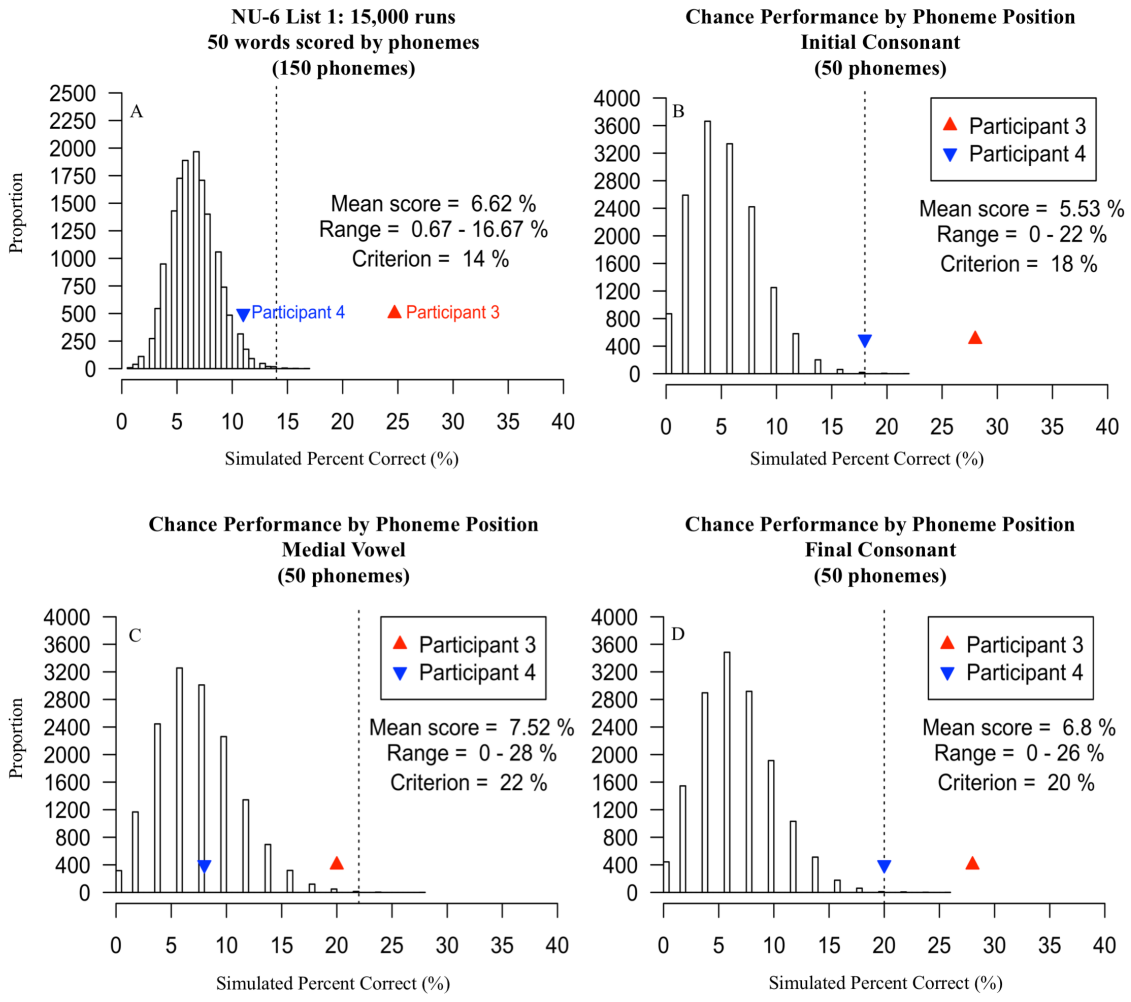

Figure 6. Effect of phoneme position on phoneme chance performance

Comparison of the simulation results to human performance data

To establish the clinical uses of the simulation findings, this study revisited the behavioral data reported in Yu & Schlauch (2018) and analyzed performance for the most adverse simulated CI listening condition (2 spectral channels in a sinewave vocoder). Table 4 shows the analyses of individual participants' performance for five participants for word scores and phoneme scores. The 99.9 percentile mark (3 standard deviations above the mean) for phoneme chance performance based on the simulation was set as the cut-off criterion for the significance test. The bold maroon figures represent phoneme scores that are significantly different from the chance performance.

Consistent with the results of Tyler, Parkinson, Woodworth, Lowder, and Gantz (1997) and Munson et al. (2003), individual variability in speech perception was also observed in normal-hearing listeners' performance in a simulated CI condition. This individual variability was observed in both word and phoneme scores for the NU-6 WR task using a single, 50-word list (Table 4). It is worth noting that Participant 5 only had one word correct (2% on the word score), which was at or near floor performance; however, the overall phoneme score and the initial consonant score of the same participant fell beyond the 99.9 percentile on the distributions of chance performance, indicating that the obtained phoneme score represented that the perception of phoneme differences was unlikely by chance. This example shows that solely relying on the word score alone for an estimate of speech perception ability can potentially underestimate a listener's ability to recognize individual speech sounds.

The following case further illustrates that the consequence of the mere dependence on a word score is possibly over- or under-estimating their ability to

perceive speech. As shown in Table 4 Participant 3 and 4 had an identical word score of 4%, meaning that both listeners had 2 words correctly identified out of 50. This score would lead clinicians to assume that both participants would have comparable skills of speech perception. However, when analyzing the WR performance by the percentage of correct phonemes as a whole (the overall phoneme score) and by phoneme position (the initial consonant, medial vowel, and final consonant scores), the results indicated that Participant 3's performance was likely to be more meaningful and less likely to be chance score than Participant 4's performance, whose scores did not show difference from the chance (Table 4). Figure 6 illustrates the relative performance of Participant 3 and 4 in relation to the obtained distributions of chance performance. These results show that a phoneme score provides a much more refined and precise method to estimate the speech perception ability than a word score, which tends overlook the contribution of individual speech sounds to a whole word recognition.

Table 4. Analyses of individual participants' performance in a simulated CI condition with 2 spectral channels

Participant	Word Score 50 items	Phoneme Score (Criterion = 14%) 150 items	Initial Consonant (Criterion = 18%) 50 items	Medial Vowel (Criterion = 22%) 50 items	Final Consonant (Criterion = 20%) 50 items
1	6%	**24.7%**	**26%**	22%	**30%**
2	20%	**44%**	**50%**	**28%**	**58%**
3	4%	**24.7%**	**28%**	20%	**28%**
4	4%	11%	18%	8%	20%
5	2%	**14.7%**	**30%**	18%	6%

Effect of list size

To quantify the effect of the list size of a WR task on phoneme chance performance, this study compared and contrasted the statistics and distributions of chance performance on the NU-6 WR task using one list (50 words) with two lists (100 words) as shown in Figure 7. Visual inspection shows the trend of decreasing the variability of chance performance as the list size doubling from 50 words to 100 words. For instance, the range of overall chance performance for a single list was 0.67 – 16.67% (Figure 6A) but was 2.33 – 12.67% for two lists (Figure 7A). Meanwhile, when using two lists of test words, the variability of chance scores was largely decreased by a factor of 1.4 (the square root of 2) from a standard deviation of 0.021 for one list to 0.014 for two lists.

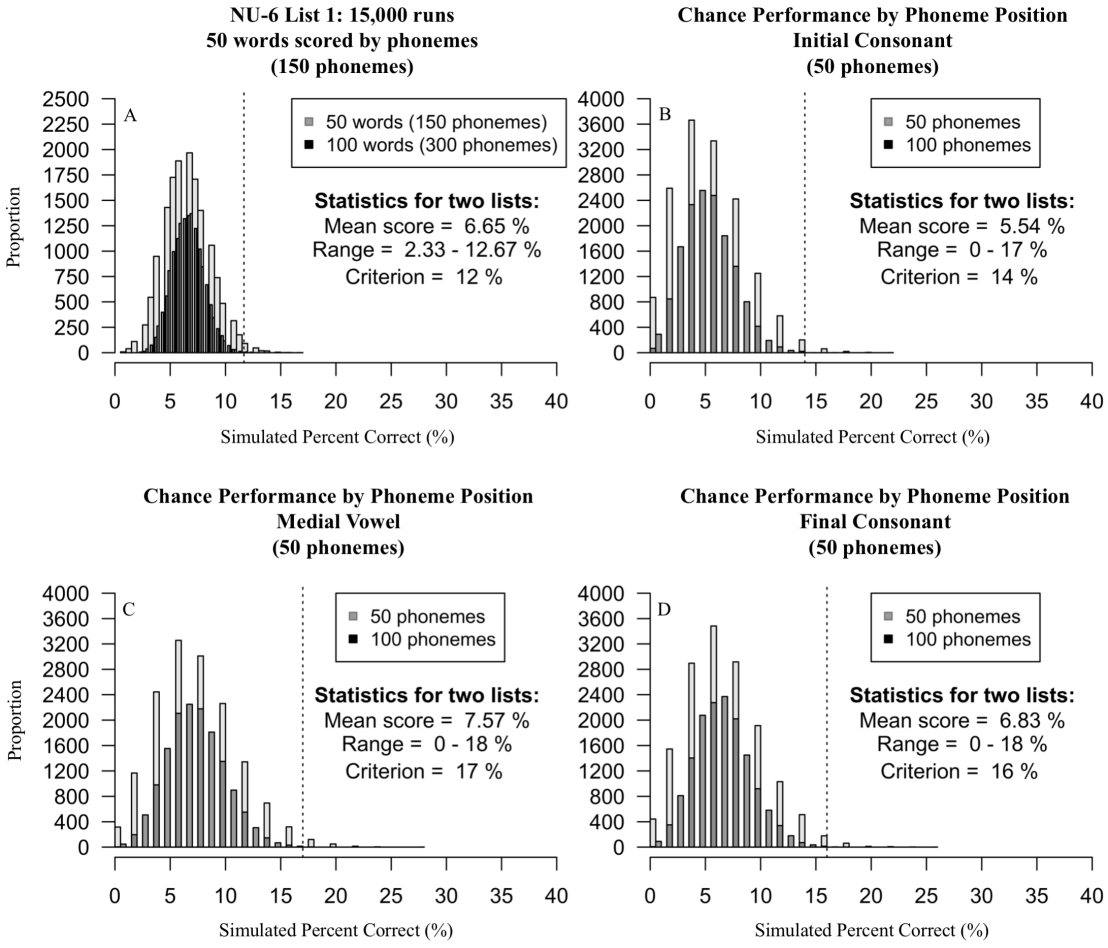

Figure 7. Effect of list size of a WR test on phoneme chance performance

Discussion

The purpose of this study was to advance the clinical utility of phoneme analysis of word scores by considering chance performance. Phoneme scores have the potential to improve the precision of diagnostic evaluations for persons with CI. Analyses were conducted using computer simulation and were applied to evaluate individual listener's scores. Findings show that estimates of chance performance for phoneme scores obtained by simulation were not affected by the choice of NU-6 list or by the word frequency of the response pool. The mean value for chance performance for phonemes is nearly identical with the predictions based on simple probability. Through simulations, this study derived the mean chance performance for each NU-6 list was 6.6%, which was consistent with a calculation suggested by A. Boothroyd (personal communication, March 3, 2018). He assumed a free impact of phonotactic constraints on listeners' random responses. Based on this assumption, the average probability of guessing a phoneme correct on a list of CNC words comprised by the 10 most frequent vowels and 20 most frequent consonants in English is 6.7% ((1/20 + 1/10 + 1/20) / 3 = 0.067).

The results of the simulation in the current study are supported also by an independent simulation performed using a different corpus of English CNC response words. This study used 2269 CNC words from the ELP (Balota et al., 2007) as the pool of random responses and derived distributions of phoneme chance scores on the NU-6 that were nearly identical with a simulation where a total of 1336 CNC words from Webster's Pocket Dictionary were used as the pool of random responses (personal communication, E. Carney, March 30, 2018). The estimated mean and range of chance performance based on the ELP corpus were 6.6% and 0.7 – 16.7%, in contrast to

Carney's simulation findings of 6.76% and 0 – 18% (personal communication, March 30, 2018).

Using word materials to assess speech recognition and analyzing the performance by percentage of correct phonemes has shown many valuable clinical implications that other types of speech recognition tests do not have. First, analyzing a WR score by phonemes is able to reveal individual differences in speech perception for persons with identical WR scores. It also provides a more refined diagnostic resolution by evaluating the performance by initial, medial, and final phonemes. This study demonstrates an example where listeners with comparable word scores (Participant 3 and 4 in Table 4 and Figure 6) can have very different abilities to perceive speech sounds that Participant 3's performance was unlikely due to chance whereas Participant 4's performance was likely due to chance. Thus, the word score itself, for a single list of 50 words, does not have enough resolution to show the individual variability in speech perception ability. In addition, studies have shown that with the same amount of time to administer the WR test analyzing WR scores by phonemes largely improved the diagnostic precision in detecting a change in hearing (Schlauch et al., 2014; Schlauch & Carney, 2018; Yu & Schlauch, 2018).

With consideration of the floor in performance, applying the phoneme analysis to WR scores also increases a WR test's ability to evaluate a wider range of performance levels in speech perception. There is a huge need for using one set of speech materials to evaluate different listening conditions and speech coding strategies for advising clients on the most appropriate device setting and the suitable rehabilitative program. For instance, for individuals with unilateral CI, combinations of different device choice and

compensatory strategies can be CI alone with and without visual cues (i.e., lip reading), as well as bimodal settings with and without visual cues, in the presence of background noise. Monosyllabic word tests have the ability to compare and contrast the performance under different listening conditions without suffering from the ceiling effects (Gifford et al., 2008; Sladen et al., 2017). This study expands the findings in Sladen et al. (2017), showing that the analysis of phonemes advances the clinical utility of WR tests. Consideration of phoneme chance performance reveals the contribution of individual speech sounds to perception.

In addition, analyzing WR scores by phonemes enables clinicians to use the established 95% confidence intervals (Schlauch & Carney, 2018) to assess whether a change in hearing occurred among different listening conditions or different visits of testing. On the other hand, the evaluation of 95% confidence intervals cannot be applied to other speech materials, specifically sentences. This is mainly because of unknown contribution of context to speech recognition, resulting in unavailability of the 95% confidence intervals for sentence materials. Having stated that, tests of speech recognition for words and sentences likely examine the same underlying processes because speech perception is a unified construct (Bilger, 1984). Studies have shown strong correlations between the performance on the CNC words and sentences (e.g., Gifford et al., 2008). This suggests that not only implementing WR tests will improve the precision of diagnostic evaluation by analyzing phoneme performance but also the WR performance can be generalized to predict the speech understanding of sentences.

This study provides the estimates of chance success in the percentage of correct phonemes in the word recognition task, which can be a potential tool to evaluate how

likely the listeners' recognition scores are by chance. However, the clinical use of these estimates is based on an important assumption, that is, a listener's chance performance on a word recognition task is a result of a random process of guessing. Future research should explore potential factors involved in guessing speech sounds in response to limited access to the acoustic information.

Conclusion

In summary, this study advances the clinical utility of WR tests, and emphasizes the importance of applying phoneme analyses to WR scores as well as the need of factoring the chance performance into analysis in order to reveal the true speech perception performance at the level of individual participants. The estimates of phoneme chance scores by phoneme position for the NU-6 were reported (Figure 6), refining the diagnostic resolution of the WR tests in revealing the individual differences in speech perception. Through a series of computer simulations and the evaluation of WR performance in a simulated CI listening condition, this study presents a finer-grained method, which is, both analyzing WR scores by phoneme and taking into account of floor performance are necessary for improving the diagnostic precision of WR tests and for examining whether the WR performance is meaningful or by chance.

Chapter 4: General Discussion and Future Direction

This dissertation provides evidence showing the fundamental differences between open- and closed-set WR tasks regarding to their diagnostic accuracy in identifying a chance in hearing. With the same number of test words, scoring in the open-set task by percentage of correct phonemes was found to be the most precise test. An analysis by phoneme errors is not possible in closed-set tasks when rhyming stimuli are used as foils because they differ by only one phoneme. The number of test words necessary for equivalent diagnostic accuracy was explored among tasks. The 4-alternative closed-set tasks are found the most inefficient task format where more than an order of magnitude in words are required to achieve the same precision as the open-set word recognition task. These findings revealed large differences in the precision of WR tests, which need to be considered when selecting a measure.

The second study of this dissertation shows a fined-grained scoring method for use in open-set WR tasks. It demonstrates that analyzing the performance by percentage of correct phonemes and considering the chance performance in phoneme recognition can provide insights into a listener's perception that are not revealed using the traditional test that computes scores based on the percentage of correct words.

To conclude, this dissertation shows the importance of considering relative diagnostic precisions among tasks, the role of chance performance in phoneme scores on a WR test, and a powerful tool using the phoneme analysis in revealing individual differences in speech perception ability.

Three directions should be pursued in future research. First, research should be done in applying the refined phoneme scoring strategy in assessing speech

understanding of clinical populations to examine the diagnostic precision of this scoring strategy in differentiating a wide range of WR performance. Second, future studies should explore the possibility of developing a closed-set task that has comparable diagnostic precision with open-set tasks. By developing an equivalent closed-set WR task, the impact of hearing loss on speech understanding for a subset of clinical population with limited speech skills can be directly and accurately assessed. Third, estimates of phoneme chance performance on a variety of clinical open-set tasks should be accomplished and tabled to empower clinicians in diagnostic evaluation of hearing loss.

Bibliography

ANSI (1997). *Methods for calculation of the speech intelligibility index* (S3.5).

ANSI (2009). *Methods for measuring the intelligibility of speech over communication systems* (S3.2).

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39*, 445-459.

Bilger, R. C. (1984). Speech recognition test development. *American Speech Hearing Association Reports, 14*, 2-15.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America, 84*(1), 101-114. doi:10.1121/1.396976

Carney, E., & Schlauch, R. S. (2007). Critical difference table for word recognition testing derived using computer simulation. *Journal of Speech Language and Hearing Research, 50*(5), 1203-1209. doi:10.1044/1092-4388(2007/084)

Chung, B. J., Hall, J. W., Buss, E., Grose, J. H., & Pillsbury, H. C. (2004). Ménière's disease: Effects of glycerol on tasks involving temporal processing. *Audiology and Neurotology, 9*(2), 115-124. doi:10.1159/000076002

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology, 17*(5), 331-349. doi:10.3766/jaaa.17.5.4

Deprez, H., Yilmaz, E., Lievens, S., & Van Hamme, H. (2013). Automating speech

reception threshold measurements using automatic speech recognition. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies* (pp. 35-40).

Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *The Journal of the Acoustical Society of America, 104*(6), 3583-3585.

Elfeky, M., Bastani, M., Velez, X., Moreno, P., & Waters, A. (2016, December). Towards acoustic model unification across dialects. *In Spoken Language Technology Workshop (SLT), 2016 IEEE* (pp. 624-628). IEEE.

Finney, D. J. (1952). *Probit analysis*. Cambridge: Cambridge University Press.

Feeney, M. P. (1990). Distinctive feature scoring of the California Consonant Test. *Journal of Speech and Hearing Disorders, 55*(2), 282-289.

Francart, T., Moonen, M., & Wouters, J. (2009). Automatic testing of speech recognition. *International Journal of Audiology, 48*(2), 80-90.

Gelfand, J. T., Christie, R. E., & Gelfand, S. A. (2014). Large-corpus phoneme and word recognition and the generality of lexical context in CVC word perception. *Journal of Speech, Language, and Hearing Research, 57*(1), 297-307.

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology and Neurotology, 13*(3), 193-205.

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *The Journal of the Acoustical Society of America, 87*(6), 2662-2674.

Halpin, C., & Sims, K. (2008). Twenty years of audiology in a patient with Norrie disease. *International Journal of Pediatric Otorhinolaryngology, 72*(11), 1705-1710. doi:10.1016/j.ijporl.2008.08.007

Han, H. J., Schlauch, R. S., & Rao, A. (2014). The effect of visual cues on scoring of clinical word-recognition tests. *American journal of audiology, 23*(4), 385-393.

Hood, J. D., & Poole, J. P. (1980). Influence of the speaker and other factors affecting speech intelligibility. Audiology, 19(5), 434-455.

Hornby, A. S., Wehmeier, S., McIntosh, C., Turnbull, J., & Ashby, M. (2005). *Oxford Advanced Learner's Dictionary of Current English 7th Edition.* New York: Oxford University Press.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed response set. *The Journal of the Acoustical Society of America, 37*(1), 158-166. doi:10.1121/1.1909295

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.

Kochkin, S. (2007) *The Impact of Untreated Hearing Loss on Household Income*. Alexandria, VA: Better Hearing Institute.

Letowski, T. R., & Scharine, A. A. (2017). *Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission* (No. ARL-TR-8227). US Army Research Laboratory Aberdeen Proving Ground United States.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19*(1), 1-36.

Lum, D. S., & Braida, L. D. (2000). Perception of speech and non-speech sounds by listeners with real and simulated sensorineural hearing loss. *Journal of Phonetics, 28*(3), 343-366.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203-208.

Martin, F. N., & Sides, D. G. (1985). Survey of current audiometric practices. *ASHA, 27*(2), 29–36.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of experimental psychology, 41*(5), 329-335.

Modified Rhyme Test Audio Library. (n.d.). Retrieved September 09, 2017, from https://www.its.bldrdoc.gov/outreach/audio/mrt_library/overview/index.htm

Munson, B., Donaldson, G. S., Allen, S. L., Collison, E. A., & Nelson, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *The Journal of the Acoustical Society of America, 113*(2), 925-935

Murphy, G. (2004). *The Big Book of Concepts*. Cambridge, MA: MIT press.

Nelson, D. A., & Chaiklin, J. B. (1970). Writedown versus talkback scoring and scoring bias in speech discrimination testing. *Journal of Speech, Language, and Hearing Research*, *13*(3), 645-654.

Neuman, A. C., Wroblewski, M., Hajicek, J., & Rubinstein, A. (2010). Combined effects

of noise and reverberation on speech recognition performance of normal-hearing children and adults. *Ear and hearing, 31*(3), 336-344.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085-1099.935.

Owens, E., & Schubert, E. D. (1968). The development of constant items for speech discrimination testing. *Journal of Speech, Language, and Hearing Research, 11*(3), 656-667.

Pavlovic, C. V. (1988). Articulation index predictions of speech intelligibility in hearing aid selection. *Asha, 30*(6/7), 63-65.

Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders, 27*(1), 62-70.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech, Language, and Hearing Research, 29*(2), 146-154.

Ross, M., & Lerman, J. A. Y. (1970). A picture identification test for hearing-impaired children. *Journal of Speech, Language, and Hearing Research, 13*(1), 44-53.

Sahu, P., Dua, M., & Kumar, A. (2018). Challenges and issues in adopting speech recognition. In *Speech and Language Processing for Human-Machine Communications* (pp. 209-215). Singapore: Springer.

Schlauch, R. S., Anderson, E. S., & Micheyl, C. (2014). A demonstration of improved precision of word recognition scores. *Journal of Speech Language and Hearing Research, 57*(2), 543. doi:10.1044/2014_jslhr-h-13-0017

Schlauch, R. S., & Carney, E. (2018). Clinical strategies for sampling word recognition performance. *Journal of Speech, Language, and Hearing Research, 61*(4), 936-944.

Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *The Journal of the Acoustical Society of America, 88*(2), 732-740.

Sladen, D. P., Gifford, R. H., Haynes, D., Kelsall, D., Benson, A., Lewis, K., ... & Westerberg, B. (2017). Evaluation of a revised indication for determining adult cochlear implant candidacy. *The Laryngoscope, 127*(10), 2368-2374.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing, 18*(2), 89.

Spahr, A. J., & Dorman, M. F. (2004). Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices. *Archives of Otolaryngology–Head & Neck Surgery, 130*(5), 624-628.

Starr, A., Picton, T. W., Sininger, Y., Hood, L. J., & Berlin, C. I. (1996). Auditory neuropathy. *Brain, 119*(3), 741-753.

Swanepoel, D. W., Clark, J. L., Koekemoer, D., Iii, J. W., Krumm, M., Ferrari, D. V., . . . Barajas, J. J. (2010). Telehealth in audiology: The need and potential to reach underserved communities. *International Journal of Audiology, 49*(3), 195-202. doi:10.3109/14992020903470783

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination*

utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6. Evanston, IL: Northwestern University, Auditory Research Laboratory.

Thornton, A. R., & Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech Language and Hearing Research, 21*(3), 507. doi:10.1044/jshr.2103.507

Tyler, R. S., Parkinson, A. J., Woodworth, G. G., Lowder, M. W., & Gantz, B. J. (1997). Performance over time of adult patients using the Ineraid or Nucleus cochlear implant. *The Journal of the Acoustical Society of America, 102*(1), 508-522.

Van Dijk, J. E., Duijndam, J. & Graamans, K. (2000). Acoustic neuroma: Deterioration of speech discrimination related to thresholds in pure-tone audiometry. *Acta Oto-Laryngologica, 120*(5), 627-632. doi:10.1080/000164800750000450

World Health Organization. (2008). *The global burden of disease: 2004 update*. Geneva: World Health Organization.

Williams, C. E., & Hecker, M. H. (1968). Relation between intelligibility scores for four test methods and three types of speech distortion. *The Journal of the Acoustical Society of America, 44*(4), 1002-1006. doi:10.1121/1.1911189

Footnote

[1]Recognition of each phoneme in a word is not independent. Boothroyd and Nittrouer (1988) found that to recognize a CNC word required an identification of only 2.5 phonemes, showing the dependency of phonemes in a word.

[2]In this study, words evaluated by phonemes for 25 words have approximately 75 phonemes. Due to lack of independence of the phonemes in a word (Boothroyd & Nittrouer, 1988), the estimates of confidence intervals were derived in Appendix C using the effective number of phonemes necessary to identify a CNC word, which is 2.5 phonemes. Therefore, Table of 95% confidence intervals for 63 phonemes (25 x 2.5 = 62.5) in Appendix C using the methods described in Carney and Schlauch (2007) were referred in the analysis of speech scores evaluated by phonemes on a 25-word task.

[3]Clopper et al. (2006) found that increased number of alternatives (from 6 to 12) and phonetic confusability together produce a more difficult closed-set task where lexical competition and talker variability effects were observed.

Appendix A

Rules for Scoring Percentage of Correct Phonemes

The following rules were modified from Schlauch et al (2014):

1. [The order Rule] If a response contained multiple phonemes, the order of the phonemes in the response had to be the same as those in the target word.
2. Regional dialects were not considered. Slight regional dialectal variations in vowels and diphthongs were considered correct if the spoken response matched the target word and not another English word.
3. [Cases of missing the consonant in the first phoneme position] If responses did not have a consonant to begin with while the target begins with a consonant, then the order rule is not applicable.
   - Example: Respond 'it' for 'dip'

     The phoneme score was 1 out of 3 because the first phoneme 'd' in the target was omitted and the last phoneme 'p' was misidentified in the response.
4. [Cases of having more phonemes in a response than a target] The order rule is not applicable when (1) a response had a consonant to begin with while the target did not, or (2) a response had an inserted phoneme.
   - Example: Respond 'feel' for 'eel'

The phoneme score was 1 out of 2 because the first phoneme 'f' in the response was an insertion

Appendix B

Computer simulation: True positive rate in detecting a change in hearing in the 50-word test condition

Step 1: Set the parameters of test lists, including the size of test list and the number of simulated scores to be generated

```
nLists = 1
nRuns = 15000*nLists
sizeList = 50
set.seed(14277)
```

Step 2: Retrieve the group-mean psychometric functions in Figure 1 to represent the underlying "true" scores for an average listener.

```
# Group means extracted from the psychometric function in the
# closed-set 4-alternative task in noise
M_N_CS4_ave <- c(82.3800, 85.8400, 90.6800, 95.9800)/100
M_N_CS4_ave_N8 <- M_N_CS4_ave[1]
M_N_CS4_ave_N4 <- M_N_CS4_ave[2]
M_N_CS4_ave_0 <- M_N_CS4_ave[3]
M_N_CS4_ave_4 <- M_N_CS4_ave[4]
```

Step 3: Generate random numbers based on a binomial model using the group mean score and the number of test items as parameters, using the closed-set 4-alternative condition as the example.

```
# rbinom(n,size,prob): generate n random numbers with
# parameters (1) number of trials (size) and (2) probability
# of success on each trial
## SNR: -8 dB
# Generate n random numbers (numbers of correct responses)
table_M_N_CS4_N8 = matrix(rbinom(nRuns, sizeList,
M_N_CS4_ave_N8), ncol=nLists)
# Convert the number of correct responses into percent score
table_M_N_CS4_N8 = table_M_N_CS4_N8/sizeList
## SNR: -4 dB
```

```
table_M_N_CS4_N4 = matrix(rbinom(nRuns, sizeList,
M_N_CS4_ave_N4), ncol=nLists)
table_M_N_CS4_N4 = table_M_N_CS4_N4/sizeList
## SNR: 0 dB
table_M_N_CS4_0 = matrix(rbinom(nRuns, sizeList,
M_N_CS4_ave_0), ncol=nLists)
table_M_N_CS4_0 = table_M_N_CS4_0/sizeList
## SNR: 4 dB
table_M_N_CS4_4 = matrix(rbinom(nRuns, sizeList,
M_N_CS4_ave_4), ncol=nLists)
table_M_N_CS4_4 = table_M_N_CS4_4/sizeList
```

Step 4: Compute the hit rates in identifying a change of hearing (e.g., from SNR of -8 to -4 dB) using the criterion of the 95th percentile for the lower distribution (i.e., set the false-positive rate to 5%). Results were shown in Figure 3B.

```
## A change of hearing from -8 to -4 dB
# Find the score located in the 95th percentile of the
# distribution of 15,000 simulated scores in -8 dB condition
criterion = quantile(table_M_N_CS4_N8, 0.95)
# Compute the rate of miss-identifying a change in hearing
# (any scores in the distribution of -4 dB overlapping with
# the distribution of -8 dB are considered as miss)
missR = sum(table_M_N_CS4_N4<=criterion)/nRuns
# Compute the hit rate: 1 - miss rate
hitR_50MNCS4_N8N4 = 1-missR
## A change of hearing from -8 to 0 dB
criterion = quantile(table_M_N_CS4_N8, 0.95)
missR = sum(table_M_N_CS4_0<=criterion)/nRuns
hitR_50MNCS4_N8P0 = 1-missR
## A change of hearing from -8 to 4 dB
criterion = quantile(table_M_N_CS4_N8, 0.95)
missR = sum(table_M_N_CS4_4<=criterion)/nRuns
hitR_50MNCS4_N8P4 = 1-missR
```

Step 5: Repeat Step 2 to 4 for other response formats, including 6-alternative closed-set task, open-set task scored by words, and open-set task scored by phonemes.

Appendix C

Table of 95% Confidence Intervals for Speech Scores Evaluated by Phonemes on a 25-word Task

Score (%)	95% Confidence Interval	Score (%)	95% Confidence Interval	Score (%)	95% Confidence Interval
0.0	0 - 5.3	34.7	20 - 52	69.3	53.3 - 84
1.3	0 - 8	36.0	21.3 - 53.3	70.7	54.7 - 84
2.7	0 - 10.7	37.3	22.7 - 54.7	72.0	56 - 85.3
4.0	0 - 13.3	38.7	22.7 - 56	73.3	57.3 - 86.7
5.3	0 - 16	40.0	24 - 57.3	74.7	58.7 - 88
6.7	1.3 - 17.3	41.3	25.3 - 58.7	76.0	60 - 88
8.0	1.3 - 20	42.7	26.7 - 60	77.3	61.3 - 89.3
9.3	2.7 - 21.3	44.0	28 - 61.3	78.7	62.7 - 90.7
10.7	2.7 - 22.7	45.3	29.3 - 62.7	80.0	65.3 - 92
12.0	4 - 25.3	46.7	30.7 - 64	81.3	66.7 - 92
13.3	4 - 26.7	48.0	32 - 65.3	82.7	68 - 93.3
14.7	5.3 - 28	49.3	33.3 - 66.7	84.0	69.3 - 94.7
16.0	5.3 - 30.7	50.7	34.7 - 66.7	85.3	72 - 94.7
17.3	6.7 - 32	52.0	34.7 - 68	86.7	73.3 - 96
18.7	8 - 33.3	53.3	36 - 69.3	88.0	74.7 - 96
20.0	8 - 34.7	54.7	37.3 - 70.7	89.3	77.3 - 97.3
21.3	9.3 - 36	56.0	38.7 - 72	90.7	78.7 - 97.3
22.7	10.7 - 38.7	57.3	40 - 73.3	92.0	80 - 98.7
24.0	12 - 40	58.7	41.3 - 74.7	93.3	82.7 - 98.7
25.3	12 - 41.3	60.0	42.7 - 76	94.7	84 - 100
26.7	13.3 - 42.7	61.3	44 - 77.3	96.0	86.7 - 100
28.0	14.7 - 44	62.7	45.3 - 77.3	97.3	89.3 - 100
29.3	16 - 45.3	64.0	46.7 - 78.7	98.7	92 - 100
30.7	16 - 46.7	65.3	49.3 - 80	100.0	94.7 - 100
32.0	17.3 - 48	66.7	50.7 - 81.3		
33.3	18.7 - 50.7	68.0	52 - 82.7		

Appendix D

Procedure of Finalizing the ELP Corpus for the Pool of Response Words

Step 1: The corpus of monosyllabic 3-phoneme words was downloaded from the ELP project (Balota et al., 2007). The size of the initial corpus was 2544 words.

Step 2: Two English-speaking research volunteers coded three phonemes for each word entry and identified 305 non-CVC words. Only CVC words were included, resulting in a total of 2239 words left (2544 – 305 = 2239).

Step 3: Two English-speaking research volunteers and the main author referred to *Oxford Advanced Learner's Dictionary* (Hornby, Wehmeier, McIntosh, Turnbull, & Ashby, 2005) to verify the accuracy of the transcription for each word. Six words were identified as incorrectly transcribed and corrected. Six words are: *landes*, *monde*, *dail*, *femmes*, *baaed*, and *y'all*.

Step 4: Thirty wh- words were manually added back to the corpus of CVC words because they were coded as 4-phoneme words in the ELP. The thirty words added to the finalized response corpus are: *whack*, *whale*, *wham*, *what*, *wheat*, *wheel*, *wheeze*, *when*, *where*, *whet*, *which*, *whiff*, *whig*, *while*, *whim*, *whin*, *whine*, *whip*, *whirl*, *whirred*, *whirrs*, *whit*, *white*, *whiz*, *whoop*, *whoosh*, *whop*, *whorl*, *why's*, *whys*. The addition of these words resulted in the finalized corpus of 2269 CVC words (2239 + 30 = 2269) for this study.

Appendix E

Computer simulation: Exploring the effect of word frequency on chance success of WR performance scored by phonemes

Step 1: Load the file of target words and determine the list of target words

```
# Load the files: as.is=TRUE parameter interpreting
# nonnumeric data as strings rather than factors
NU6 <- read.csv ('FinalCorpus_NU6.csv', header=TRUE,
as.is=TRUE)
# Determine the list of target words in the NU6 corpus
# Randomize the order of the NU6 words and extract only
# words matching the parameter
NU6_L1 <- filter(NU6[sample(nrow(NU6), nrow(NU6)),], List ==
1)
# Extract the data frame of phonemes
NU6_L1_Phoneme <- select (NU6_L1, starts_with('Phoneme'))
```

Step 2: Load the file of response words and remove the words that are less frequent than the average word frequency of the target words

```
# Load the files
Corpus <- read.csv ('FinalCorpus_ELP.csv', header=TRUE,
as.is=TRUE)
# Finalize ELP corpus: include Only words in certain frequency
# range
NU6_L1 <- subset (NU6, List==1)
mean(NU6_L1$Log_Freq_HAL)
Corpus_f9.68 <- subset (Corpus, Log_Freq_HAL>=9.68)
```

Step 3: Set a loop operation for (1) randomly retrieve, without replacement, words from the finalized ELP corpus as response words, (2) pair each random response word with a target word from the NU-6 accordingly, and (3) compute the number of correct responses.

```
# Set the number of simulated scores representing individuals'
# performance
NumSimIndPerf = 15000
```

```
# Create a double-precision vector to store values
vals_All3Phonemes <- numeric(NumSimIndPerf)
vals_Word <- numeric(NumSimIndPerf)
# Start the comparison between two strings (one from ELP,
# another from NU6) each time, for a total of 15000 times
for (i in 1:NumSimIndPerf) {
  # Get a sample of 50 words from ELP corpus: nrow(): Get the
  # length of the ELP corpus
  smpl <- Corpus_f9.68[sample(1:nrow(Corpus_f9.68),50),]
  smpl_Word <- select(smpl, starts_with('Pron_Finalize'))
  smpl_Phoneme <- select(smpl, starts_with('Phoneme'))
  # Use apply (x, 1, sum) to sum across rows; 1: rows (overall
  # phoneme performance); 2: columns (phoneme performance at
  # each position)
  # sum() across all the rows to count number correct for the
  # 50 rows
  # The number of correct phonemes out of 150 phonemes
  vals_All3Phonemes[i] <-
sum(apply(smpl_Phoneme==NU6_L1_Phoneme, 1, sum))
  # The number of correct words out of 50 words
  vals_Word[i] <- apply(smpl_Word==NU6_L1_Word, 2, sum)
}
# Compute the percent correct phoneme and word scores
vals.pct_All3Phonemes <-
round((vals_All3Phonemes/(3*(nrow(NU6_L1)))),4)
vals.pct_Word <- vals_Word/nrow(NU6_L1)
# Save the 15,000 phoneme scores into a txt file
write.table(data.frame(vals.pct_All3Phonemes),
"SFreq_L1_vals.pct_All3Phonemes.txt", sep = "\t")
```

Step 4: Identify the cut-off score on the distribution of 15,000 scores by searching the score at the three standard deviations above the mean and set it as the criterion for the significance test.

```
# Compute the score that is at the 3 SDs above the mean as the
# criterion (3 SDs above the mean = 99.9 percentile rank)
criterion = quantile(vals.pct_All3Phonemes, 0.999)
```